Today’s Take-away

• Stakeholders and implementations
• Current and new features
• NFS community resources
• Open discussion
What is NFS/RDMA?

- NFS on a low latency copy offload transport
- RDMA replaces sockets, TCP, IP under RPC
- No impact on performance of underlying persistent storage
What Fabrics?

- InfiniBand
- RoCE (v1, v2)
- iWARP
Why Linux NFS/RDMA
Storage on RDMA

- NFS/RDMA
- iSER
- SRP
- SMB Direct
Trends

• More virtualization
  • Private: OpenStack, Exadata
  • Public: AWS, Google Cloud
• More unstructured block storage on NFS
Trends

- Persistent storage latencies going down
- Think DRAM speeds
- Storage fabric latencies have to keep up
Customers

- Low latency required
- HPC, Labs
- Cloud back-end storage
- Fabric already present
- Engineered systems
- Data center
Linux Differentiators

- Market-leading NFS client implementation
- Penetration of HPC market
- Diversity of physical file systems
- iWARP / RoCE with NFS/RDMA
Reaches Link Speed

12 readers IOzone Throughput
mount wrize,rsize=256K
Low CPU Utilization

12 readers IOzone CPU utilization

mount wsize, rsize=256K
Community Snapshot

- Individuals
- Implementations
- Stakeholders
Coming Implementations

• Ganesha server
• VMware NFSv4.1 client
• Others?
Known Implementations

• Linux client and server
• Solaris client and server
• GlusterFS server (NFSv3)
Break

Back in 10 minutes
Enterprise Linux
EL Use Cases

- GlusterFS
- Ganesha
- OpenStack Cinder
- RHS
- Others?
Upstream Client Plans

• NFSv4.1 / pNFS
• Small I/O performance
• Scalability (NUMA, many mounts)
• High availability environments
• Adaptor hot-plug
Upstream kNFSD Plans

- kNFSD is a reference implementation
- Still missing a full-time subsystem maintainer
Troubleshooting Challenges

- ibdump - mlx4 only
- Wireshark - no RPC/RDMA dissector
- rpcdebug - known limitations
Enabling Full Support

- Q/A resources
- Hardware
- Engineering
- Community support
- Adapter diversity
Break

Back in 10 minutes
Community Issues
Continuous Testing

- Functional tests
  - cthon04, xfstests
- Performance / stress
  - iozone, fio, dbench
Community Testing Events

- Are we ready for NFS/RDMA plug-fests?
- Infrastructure requirements: What fabrics?
- Additional testing events?
- New test software?
Protocol Enhancements?

- NFSv4.1
- Backchannel
- Credit limit and session slot table size
- pNFS
Protocol Enhancements?

- Capability management
- Inline buffer sizes
- Server remote invalidation
- Multiple QPs per transport
Protocol Enhancements?

- Multiple payloads per RPC
- Faster bring-up of new implementations
Open Discussion

The Tirpitz, or Enigma-T, is a non-standard Enigma especially designed for use in communications between Germany and Japan during WWII. This machine is a variant of the Enigma-K and was supplied with eight rotors, each with five turn-over notches.

The Tirpitz was called “TIRPITZ” by the Germans, and spelled as “TIRUPYITSU” by the Japanese.
Appendix